**DATA MINING PROJECT DOCUMENT**

***House Prices***

***Advanced Regression Techniques using XGboost***



**Disusun oleh:**

|  |  |
| --- | --- |
| 12S21009 | Mikhael Janugrah Pakpahan |
| 12S21010 | Bobby Willy Siagian |
| 12S21011 | Aldi Jeremy Simamora |

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO**

**INSTITUT TEKNOLOGI DEL**

**2024/2025**

**DAFTAR ISI**

[DAFTAR GAMBAR 4](#_Toc184590353)

[DAFTAR TABEL 5](#_Toc184590354)

[BAB 1 BUSINESS UNDERSTANDING 6](#_Toc184590355)

[1.1 Determine Business Objective 6](#_Toc184590356)

[1.2 Determine Project Goal 7](#_Toc184590357)

[1.3 Produce Project Plan 7](#_Toc184590358)

[BAB 2 DATA UNDERSTANDING 9](#_Toc184590359)

[2.1 Collecting Data 9](#_Toc184590360)

[2.2 Describe Data 9](#_Toc184590361)

[2.3 Data Validation 12](#_Toc184590362)

[BAB 3 DATA PREPARATION 13](#_Toc184590363)

[3.1 Data Selection 13](#_Toc184590364)

[3.2 Data Cleaning 14](#_Toc184590365)

[3.3 DATA TRansformation 15](#_Toc184590366)

[BAB 4 MODELLING 17](#_Toc184590367)

[4.1 Building Test Scenario 17](#_Toc184590368)

[4.2 Built Model 18](#_Toc184590369)

[BAB 5 EVALUATION 20](#_Toc184590370)

[5.1 Evaluate Results 20](#_Toc184590371)

[5.2 Evaluate Process 20](#_Toc184590372)

[BAB 6 DEPLOYMENT 21](#_Toc184590373)

[DAFTAR PUSTAKA 22](#_Toc184590374)

# DAFTAR GAMBAR

[Gambar 1. Collecting Data 10](#_Toc184590279)

[Gambar 2. Train data head 13](#_Toc184590280)

[Gambar 3. Sample Submission Data Head 13](#_Toc184590281)

[Gambar 4. Data Selection 14](#_Toc184590282)

[Gambar 5. Missing value pada file fktpkapitasi.dta 15](#_Toc184590283)

[Gambar 6. Data Transformation 17](#_Toc184590284)

[Gambar 7. Project Completion Methodology Flowchart 18](#_Toc184590285)

[Gambar 8. Built Model XGboost 19](#_Toc184590286)

[Gambar 9. Evaluating prediction 20](#_Toc184590287)

# dAFTAR TABEL

[Tabel 1. Project Plan 7](#_Toc184590342)

[Tabel 2.Variable Information 10](#_Toc184590343)

# BAB 1 BUSINESS UNDERSTANDING

Pada bab ini akan dibahas tentang penentuan objektif bisnis, tujuan bisnis, dan membuat rencana proyek.

## 1.1 Determine Business Objective

Dalam konteks analisis harga rumah, pemahaman bisnis (business understanding) menjadi langkah pertama yang sangat penting untuk membangun solusi yang tepat guna. Penentuan harga rumah dalam industri properti melibatkan berbagai faktor yang kompleks dan dinamis, yang tidak hanya berkaitan dengan nilai ekonomi dari sebuah properti, tetapi juga dipengaruhi oleh kondisi pasar, lokasi, karakteristik fisik rumah, serta faktor lingkungan dan sosial yang lebih luas. Bagi agen properti, pemilik rumah, investor, dan calon pembeli, harga rumah menjadi salah satu parameter utama yang digunakan untuk pengambilan keputusan. Oleh karena itu, adanya metode yang dapat memprediksi harga rumah dengan akurasi yang lebih tinggi sangat penting untuk meningkatkan efisiensi dan mendukung keputusan yang lebih cerdas.

Secara tradisional, penentuan harga rumah sering kali mengandalkan analisis manual yang dilakukan oleh agen atau evaluator properti. Proses ini biasanya memerlukan pengetahuan pasar lokal yang mendalam dan pemahaman terhadap berbagai faktor yang mempengaruhi harga, yang sering kali bersifat subjektif dan rentan terhadap kesalahan. Selain itu, fluktuasi pasar yang dipengaruhi oleh kondisi ekonomi, kebijakan pemerintah, dan tren pasar yang terus berubah, semakin menambah kompleksitas dalam proses ini.

Untuk mengatasi tantangan ini, pendekatan berbasis machine learning menawarkan solusi yang lebih otomatis dan akurat. Dalam hal ini, algoritma XGBoost (eXtreme Gradient Boosting) menjadi pilihan yang sangat tepat untuk digunakan dalam prediksi harga rumah. XGBoost adalah metode pembelajaran mesin yang menggunakan teknik boosting, di mana sejumlah model pohon keputusan yang relatif sederhana (weak learners) digabungkan untuk membentuk sebuah model yang kuat dan akurat. Dengan kemampuan untuk menangani masalah regresi dengan baik, XGBoost mampu memprediksi harga rumah dengan lebih presisi, memperbaiki kesalahan prediksi secara iteratif, dan memberikan hasil yang lebih stabil serta konsisten dibandingkan dengan metode konvensional.

Pemahaman bisnis dalam proyek ini adalah untuk memastikan bahwa model prediksi harga rumah tidak hanya memberikan akurasi tinggi, tetapi juga dapat memberikan wawasan yang dapat dipahami oleh pemangku kepentingan industri properti. Dengan demikian, proyek ini tidak hanya bertujuan untuk menciptakan solusi teknis yang efisien, tetapi juga untuk memastikan bahwa hasil prediksi yang diberikan relevan, transparan, dan dapat digunakan untuk mendukung keputusan strategis di pasar properti yang sangat dinamis.

## 1.2 Determine Project Goal

Tujuan utama dari proyek ini adalah mengembangkan model prediksi menggunakan algoritma XGBoost (eXtreme Gradient Boosting), yang terbukti efektif untuk masalah regresi terutama pada data terstruktur seperti dataset harga rumah. Model ini akan dilatih dengan data historis tentang properti dan harga jualnya, dengan sasaran untuk menghasilkan prediksi harga yang akurat untuk properti baru yang belum pernah muncul dalam data sebelumnya. Teknik XGBoost dipilih karena kemampuannya menangani fitur kompleks dan keandalannya dalam meningkatkan akurasi prediksi melalui metode boosting.

## 1.3 Produce Project Plan

Berikut adalah jadwal pengerjaan proyek yang akan dilakukan:

Tabel 1. Project Plan

|  |  |  |
| --- | --- | --- |
| **Tahapan** | **Waktu Pengerjaan** | **Kegiatan** |
| *Business Understanding* | 3 hari | Pada tahap ini akan dilakukan penentuan terhadap objektif bisnis, menentukan tujuan proyek, dan membuat rencana proyek. |
| *Data Understanding* | 3 hari | Pada tahap ini data yang akan digunakan akan dikumpulkan, dianalisis, dan divalidasi. |
| *Data Preparation* | 4 hari | Pada tahap ini akan dilakukan *data* cleaning, *data construction*, dan *data integration*. |
| *Modeling* | 3 hari | Pada tahap ini akan dilakukan pemodelan terhadap model serta pengujian model. |
| *Evaluation* | 3 hari | Pada tahap ini akan dilakukan evaluasi terhadap hasil model yang dibangun dan melakukan review terhadap proses pembangunan model. |
| *Deployment* | 4 hari | Pada tahap ini akan dilakukan *deployment*, *monitoring*¸dan *maintenance* terhadap model. |

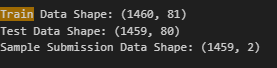
Dalam pelaksanaan proyek ini, Python menjadi salah satu alat penting yang akan digunakan. Python adalah bahasa pemrograman serbaguna yang mampu menjalankan berbagai perintah secara langsung melalui metode berorientasi objek dan menggunakan sintaks yang mudah dibaca. Python terkenal dalam bidang Data Science, Machine Learning, dan Internet of Things (IoT), menjadikannya pilihan utama untuk analisis data dan pembuatan model prediksi. Pada proyek ini, algoritma utama yang digunakan adalah XGBoost (Extreme Gradient Boosting). XGBoost merupakan algoritma machine learning yang menggunakan pendekatan ensemble learning, mirip dengan metode Random Forest. Namun, XGBoost memiliki fokus yang berbeda, yaitu menghasilkan model pohon keputusan secara bertahap untuk meningkatkan kinerja dan akurasi prediksi. Dalam XGBoost, setiap pohon (tree) dibangun secara berurutan, dengan bobot khusus yang diberikan pada masing-masing pohon untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. Dengan strategi ini, XGBoost mampu menangkap pola yang lebih kompleks dalam data, membuatnya sangat efektif untuk prediksi harga rumah.

# BAB 2 DATA UNDERSTANDING

## 2.1 Collecting Data

Dalam tahapan *data understanding*, yang merupakan langkah awal untuk memahami data yang akan digunakan, proses ini dimulai dengan pengumpulan, deskripsi, dan pemahaman data yang relevan untuk penelitian. Dataset yang digunakan dalam penelitian ini terdiri dari tiga file utama, yaitu train.csv, test.csv, dan sample\_submission.csv. Dataset ini digunakan untuk memprediksi harga rumah berdasarkan karakteristik properti yang telah disediakan.

Dalam tahapan data understanding yang merupakan tahapan pemahaman terhadap data yang akan digunakan, tahapan ini dimulai dari mengumpulkan data, mendeskripsikan data dan memahami data yang akan digunakan dalam penelitian.



Gambar 1. Collecting Data

## 2.2 Describe Data

Pengumpulan data merupakan tahap awal untuk menemukan data yang akan digunakan dalam penelitian. Maka dari itu dataset yang akan digunakan untuk memprediksi harga rumah berdasarkan karakteristik properti yang telah disediakan berdasarkan dataset *train, test, dan* sample\_submission .

* Sheet test adalah sheet untuk untuk menguji model prediksi. Dataset ini berisi fitur properti tanpa nilai harga rumah (*target variable*).
* Sheet train berisi data historis properti, termasuk fitur-fitur seperti ukuran properti, jumlah kamar, lokasi, dan harga rumah (*target variable*).
* Sheet sample\_submission berisi template untuk mengunggah hasil prediksi harga rumah dari dataset test.csv. Dimana format kolomnya biasanya mencakup ID properti dan kolom prediksi harga rumah.

Berikut ini adalah variable yang terdapat pada set data untuk memprediksi harga rumah berdasarkan karakteristik properti yang telah disediakan.

Tabel 2.Variable Information

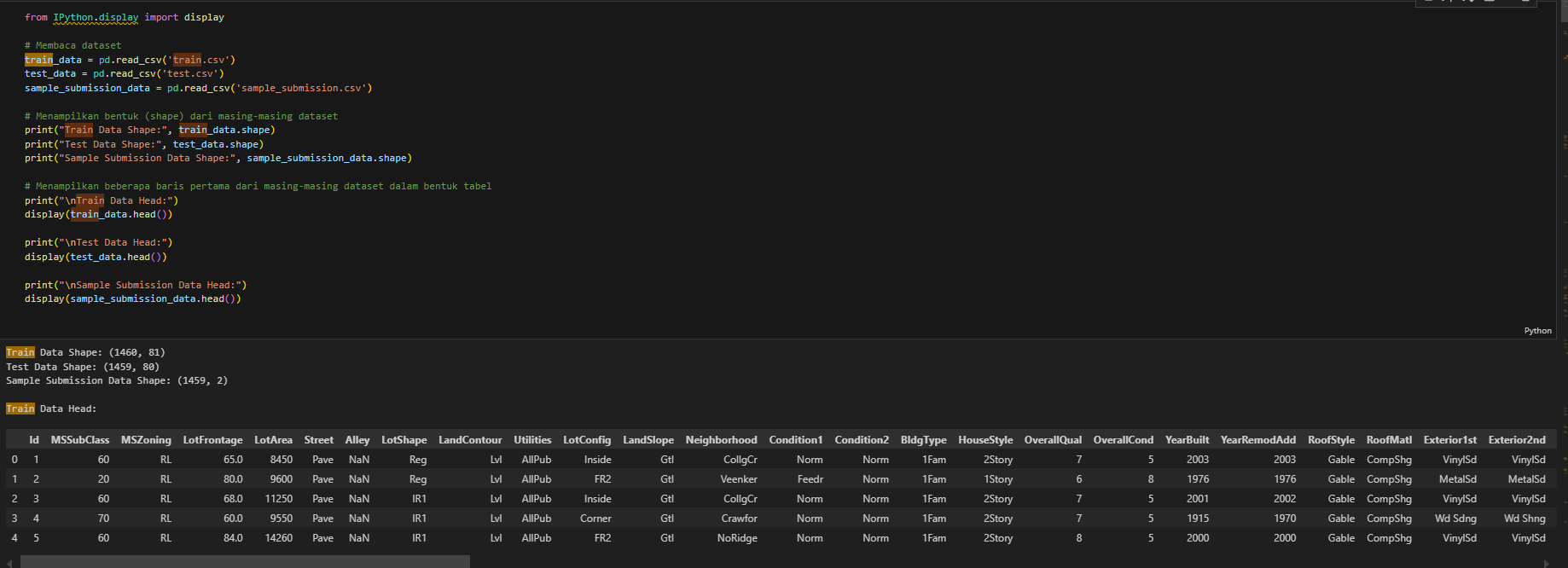
| **Variable Information** | | |
| --- | --- | --- |
| **Variable** | **Position** | **Label** |
| Id | 1 | |  | | --- | | Identifikasi unik untuk setiap properti | |
| MSSubClass | 2 | Tipe bangunan tempat tinggal (kategori kode numerik) |
| MSZoning | 3 | |  | | --- | | Zona perumahan (kategori) | |
| LotFrontage | 4 | |  | | --- | | Panjang properti ke jalan (dalam kaki) | |
| LotArea | 5 | |  | | --- | | Luas properti (dalam kaki persegi) | |
| Neighborhood | 6 | |  | | --- | | Lokasi properti berdasarkan nama lingkungan | |
| YearBuilt | 7 | |  | | --- | | Tahun konstruksi selesai | |
| OverallQual | 8 | |  | | --- | | Kualitas material dan penyelesaian rumah (skala ordinal 1-10) | |
| OverallCond | 9 | |  | | --- | | Kondisi keseluruhan rumah (skala ordinal 1-10) | |
| TotalBsmtSF | 10 | |  | | --- | | Luas total ruang bawah tanah (dalam kaki persegi) | |
| GrLivArea | 11 | |  | | --- | | Luas ruang lantai di atas tanah (dalam kaki persegi) | |
| GarageCars | 12 | |  | | --- | | Kapasitas garasi (dalam jumlah mobil) | |
| GarageArea | 13 | |  | | --- | | Luas garasi (dalam kaki persegi) | |
| MoSold | 14 | |  | | --- | | Bulan properti terjual | |
| YrSold | 15 | |  | | --- | | Tahun properti terjual | |
| SalePrice | 16 | Harga jual properti (variabel target) |

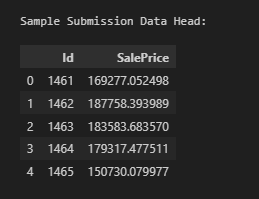
Tujuan dari proyek ini adalah membangun model XGBoost (eXtreme Gradient Boosting) untuk memprediksi harga rumah berdasarkan fitur-fitur properti, seperti ukuran tanah, lokasi, kondisi bangunan, dan fitur lainnya. Model ini akan membantu mengklasifikasikan atau memperkirakan nilai rumah berdasarkan data yang tersedia, dengan fokus pada pemanfaatan variabel seperti LotArea, Neighborhood, YearBuilt, dan OverallQual untuk memberikan estimasi harga properti (SalePrice) secara akurat.

## 2.3 Data Validation

Pada tahap ini dilakukan validasi terhadap data yang akan digunakan dengan memeriksa kelengkapan data untuk menghindari terjadinya *error* ataupun masalah *input data* yang terjadi *missing value*. Maka dari itu berikut pemeriksaan terhadap atribut utama yang akan digunakan pada dataset.

* Train Data Head

 Gambar 2. Train data head

]

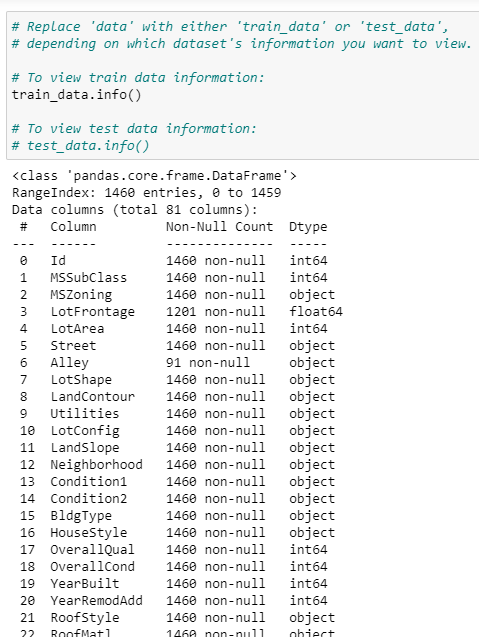
Gambar 3. Sample Submission Data Head

# BAB 3 DATA PREPARATION

Bab ini membahas tentang *data selection, data cleaning, data construct, labeling data,* dan *data integration.*

## 3.1 Data Selection

Data selection adalah proses pemilihan data yang relevan untuk digunakan dalam pengerjaan  
proyek. Pada proyek kali ini data yang di gunakan berasal dari dataset yang sudah di tentukan terlebih dahulu. Oleh karena itu dilakukan terlebih dahulu menampilkan ringkasan dataset dan memeriksa struktur dataset. Hasil data selection dapat dilihat dari gambar 4 berikut.

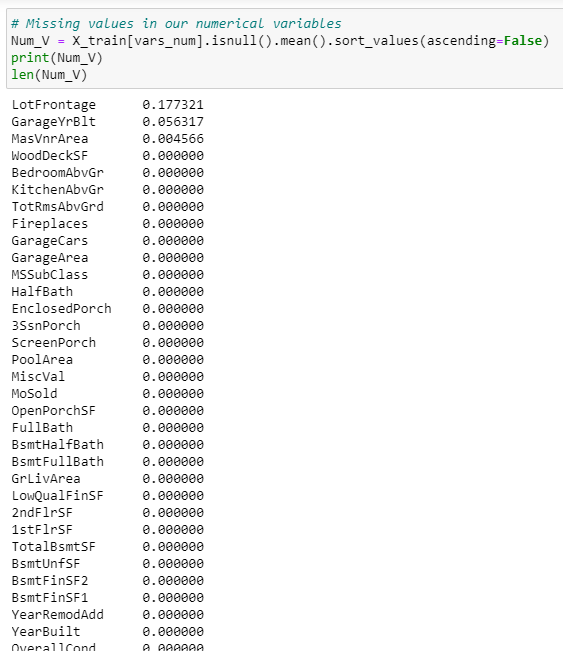


Gambar 4. Data Selection

## 3.2 Data Cleaning

Pada proyek ini, data cleaning tidak perlu dilakukan, karena data yang diberikan suda bersih.

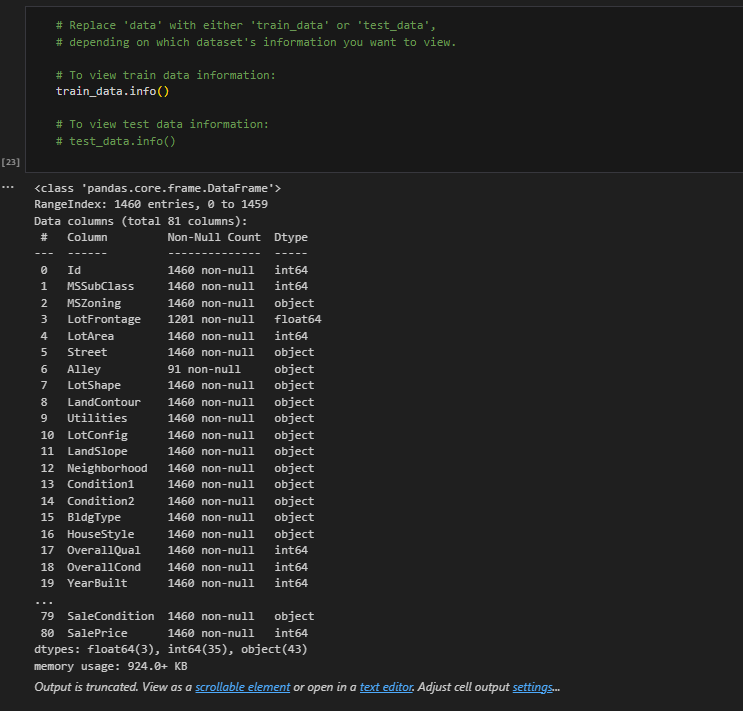
Pencarian missing value pada file fktpkapitasi.dta dapat dilihat pada gambar 5.



Gambar 5. Missing value pada file fktpkapitasi.dta

## 3.3 DATA TRansformation

Subbab ini menjelaskan transformasi data yang dilakukan dengan memahami struktur dataset melalui fungsi train\_data.info(), yang menampilkan informasi tentang 81 atribut, termasuk tipe data (int64, float64, object) dan kelengkapan nilai. Dataset memiliki 1460 baris, dengan beberapa atribut tidak lengkap, seperti LotFrontage (1201 nilai) dan Alley (91 nilai). Transformasi dilakukan dengan memilih atribut relevan, seperti OverallQual, YearBuilt, dan SalePrice untuk analisis, menangani nilai yang hilang pada atribut seperti Alley, serta menyesuaikan tipe data agar sesuai untuk analisis lebih lanjut. Transformasi ini memastikan dataset bersih, lengkap, dan relevan untuk mendukung tujuan proyek, dapat dilihat pada gambar 6.



Gambar 6. Data Transformation

# BAB 4 MODELLING

Pada bab ini, akan dijelaskan tentang model yang akan diimplementasikan pada proyek. Bab ini berisikan penjelasan tentang langkah-langkah yang akan dilakukan untuk menyelesaikan proyek. Flowchart metodologi penyelesaian proyek dapat dilihat pada Gambar 7.



Gambar 7. Project Completion Methodology Flowchart

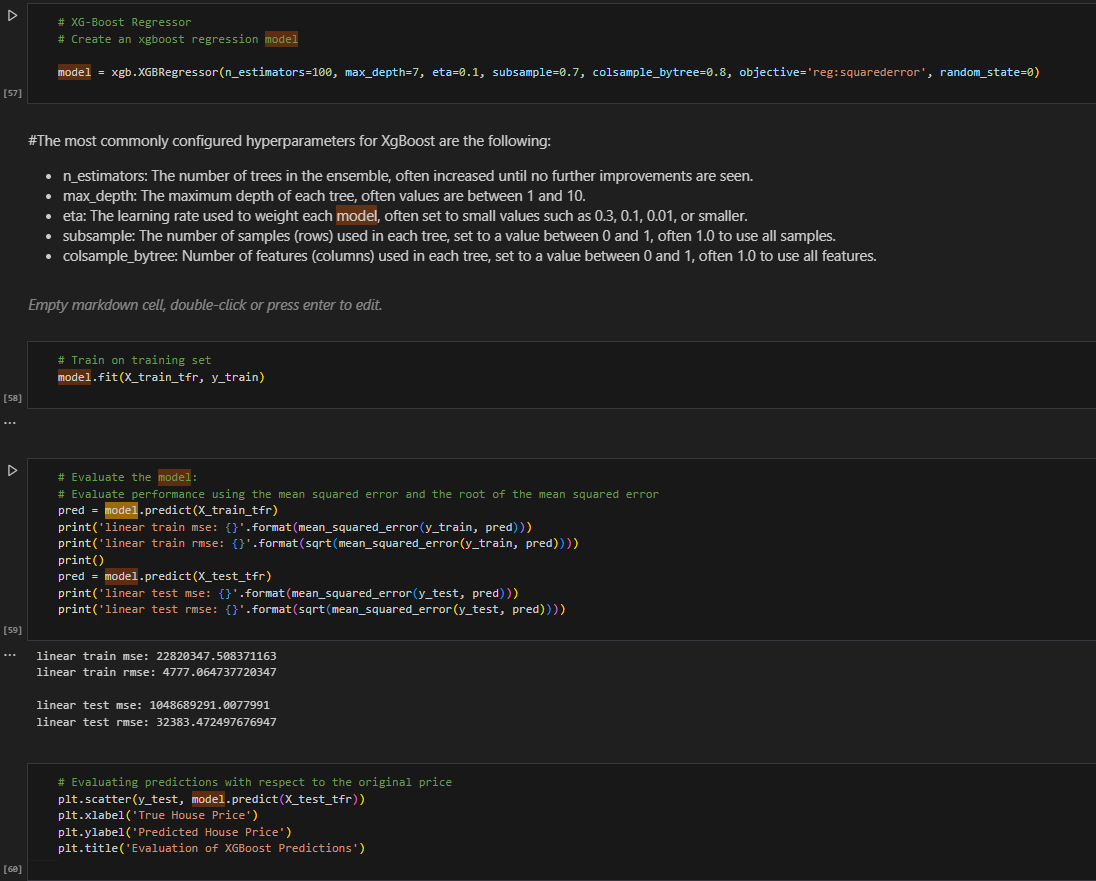
Proyek ini dimulai dengan mengumpulkan data yang akan digunakan untuk memprediksi harga rumah (*house prices prediction*). Data yang telah dikumpulkan akan melalui tahap *preprocessing* untuk memastikan kebersihan dan kelengkapan data. Setelah data diproses dan siap digunakan, tahap implementasi dilakukan dengan membangun model prediksi menggunakan algoritma XGBoost. Model ini akan memprediksi harga rumah berdasarkan berbagai atribut seperti ukuran rumah, lokasi, dan kondisi bangunan. Selanjutnya, evaluasi dilakukan untuk mengukur kinerja model menggunakan metrik evaluasi yang relevan. Tahap terakhir adalah *deployment*, di mana model yang telah dibangun diterapkan pada data uji untuk memprediksi harga rumah dalam konteks nyata.

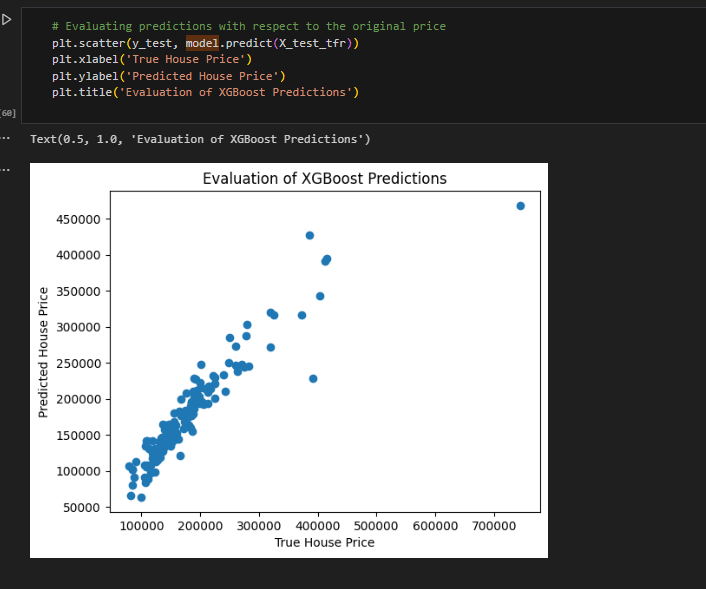
## 4.1 Building Test Scenario

Proyek ini akan menggunakan XGBoost untuk membantu memprediksi harga rumah berdasarkan berbagai atribut penting. Data yang digunakan berasal dari dataset prediksi harga rumah, dengan atribut yang relevan, seperti ukuran rumah, tahun dibangun, lokasi, serta kondisi keseluruhan rumah. Beberapa atribut utama yang akan digunakan dalam model ini adalah OverallQual (kualitas keseluruhan), YearBuilt (tahun pembangunan), dan SalePrice (harga penjualan sebagai target). Proses prediksi akan dilakukan dengan membangun skenario pengujian, seperti mengukur akurasi model dalam memprediksi harga rumah berdasarkan data uji, serta mengevaluasi performa menggunakan metrik seperti *Root Mean Squared Error* (RMSE) dan *Mean Absolute Error* (MAE) untuk memastikan model mampu memberikan prediksi harga rumah yang akurat dan andal.

## 4.2 Built Model

Setelah set data yang dikumpulkan sudah benar-benar bersih dan siap digunakan, maka tahap ini model akan dibangun dengan menggunakan data test. Data test ini diambil dari set data yang digunakan, dapat dilihat pada gambar 8.

Gambar 8. Built Model XGboost

Gambar 9. Evaluating prediction

# BAB 5 EVALUATION

## 5.1 Evaluate Results

## 5.2 Evaluate Process

# BAB 6 DEPLOYMENT

# DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [1] | B. N. S. Aji Primajaya, "Random Forest Algorithm for Prediction of Precipitation," *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM) ,* vol. Vol. 1, pp. 3-4, March 2018. |
| [2] | A. N. Rachmi, "IMPLEMENTASI METODE RANDOM FOREST DAN," *XGBOOST PADA KLASIFIKASI CUSTOMER CHURN,* p. 33, 2020. |
| [3] | D. David, "https://www.freecodecamp.org/," CodeCampNews, 06 August 2020. [Online]. Available: https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/. [Accessed tuesday 2021 2021]. |